

The Extent of Association of Copy Number Variation with Haplotype Structure at the AMY1 Locus.

Edmund Henry Gilbert¹

¹The University of Nottingham
University Park
Nottingham, NG7 2RD

mbydeg@nottingham.ac.uk

Keywords

Copy Number Variation, Haplotypes, Association, Amylase, Microsatellite, RFLP

Word Count: 5 814

Abstract

Copy number variable (CNV) regions often have phenotypic consequences, due to gene dosage effects. These consequences are thought to expose these regions to selective pressures. Previous literature has identified the CNV locus of AMY1, a gene encoding the salivary amylase isoform, appears to have been under selection (**Perry et al, 2007**); it is a locus with substantial variation with diploid total ranges from 2--15.

Haplotype blocks are regions of low diversity commonly having 2-4 major haplotypes that are inherited as discrete units. This discrete nature allows them to be used to explore population structure, and could be used to tag AMY1 CNV identity for population studies. Therefore it is necessary to ascertain whether haplotype blocks and CNV at the AMY1 locus are associated.

In order to carry out this study we designed genotyping assays for three SNPs; rs12130703, rs12075086 and rs1185098. These segregated 370 European diploid samples into 4 common haplotypes and 1 rare haplotype. 2 of these haplotypes represent 80% of the population studied. We then compared these haplotypes to CNV. This CNV was genotyped by a tetra-repeat microsatellite found in every copy of the AMY1 copy unit. This microsatellite's length polymorphism provided an extra dimension with which to describe CNV with.

We observed no significant association of CNV identity with haplotype structure at the AMY1 locus. Therefore we conclude that there is strong CNV-haplotype structure within European individuals, and haplotype identity cannot be used to predict AMY1 CNV, nor tag its identity within the population.

Introduction

Genomes, including the human genome, are not a sequence of completely independent loci, but are structured sequences with extensive association with nearby loci, so that one allele at one position can be strongly linked with another. Linkage disequilibrium (LD) describes this non-random association of loci (**Wall and Pritchard, 2003**), where alleles segregate together. This state has been described extensively (**Lewontin and Kojana, 1960; Slatkin, 2008**). LD is commonly measured using r^2 (the extent of correlation) or using D' (**Hedrick, 1987**). This 'linkage' is different from genetic linkage, which describes the tendency of closely positioned loci on a chromosome to be inherited together during meiosis.

LD is structured into blocks of associated loci, such as single nucleotide polymorphisms (SNPs) (**Daly et al, 2001**), that are present across the entire human genome (**Gabriel et al, 2002**), and present limited diversity (**Patil et al, 2001**), often having 2-4 common haplotypes per block (**Daly et al, 2001**). These blocks are flanked by recombination hotspots (**Jeffreys et al, 2001**) (**Philips et al, 2003**), therefore the contents of these LD blocks do not frequently recombine. The alleles within subsequently are not then randomised by recombination. Therefore the content of these blocks are inherited through generations with discrete identities.

In evolutionary terms these blocks evolve and mutate, creating 'branches' off of root haplotypes (**Underhill et al, 2000**). Therefore investigating these changes by mapping the diversity within a population can provide us with information on population structure and history (**Cuciani et al, 2011; Kayser et al, 2001; Zegura et al, 2003**).

Haplotype blocks are commonly investigated using SNPs, and because of the tightly associated nature of the SNPs within-block, it is only necessary to genotype a few SNPs in order to 'tag' the identity of the block as a whole (**Johnson et al, 2001**). Further analysis has shown this tagging method does not significantly hamper the statistical power of studies investigating haplotype block diversity (**Zhang et al, 2002; Zhang et al 2004**).

These haplotype blocks can also be indicative of evolutionary pressures such as a selective sweep as well as help investigating population structure (**Slatkin, 2009; Jakobsson et al, 2008**). These infrequently recombining haplotypes could be carried along with the selected locus by positive selection via linkage. This model is often called genetic 'hitchhiking' (**Smith and Haigh, 1974**). Therefore the investigation of haplotypes may allow us to study regions thought to have been exposed to selective pressures.

One class of genetic variation that is of interest to this kind of study is copy number variation (CNV); where a genomic sequence larger than 1kb presents variable copy numbers (**Conrad et al, 2010**). CNV can range from single site deletion of sequence (**McCarroll et al, 2008**) to tandem repeats of a unit. There are many examples of this class of polymorphism in the human genome. Examples include the; FcGR3 (**Koene et al, 1998; Breunis et al, 2009**), beta-defensin (**Holox et al 2003**), opsin (**Nathans et al, 1986**), olfactory (**Rouquier et al, 1998**) and the salivary amylase (**Groot et al, 1989**) loci.

Segmental duplications are a common form of CNV (**Bailey et al, 2008**) and the generation of CNV can be somatic or meiotic (**Bruder et al, 2008**). CNV can be produced by Homologous Recombination and Non-Homologous Repair Mechanisms (**Hastings et al, 2009**). There is also evidence that retrotransposons such as the Alu transposon (which is primate specific) have a role within CNV generation (**Bailey et al, 2001; Kidd et al, 2008**).

There are several methods for measuring this variation. A common form is to use a SNP array comparative-gene-hybridisation (array-CGH) method to measure the relative differences of individuals against a known reference sample (**Redon et al, 2006**). Fluorescent in situ hybridisation (FISH) techniques also give a clear visual confirmation of such measurements (**Perry et al, 2007**). More recently the paralogue ratio test (PRT) (**Armour et al, 2007**) has been used. This is a method that simultaneously amplifies the CNV locus and a single copy reference to provide a copy number (CN).

Several examples of CNV also demonstrate phenotypic consequences arising from their variation; such as salivary amylase (**Perry et al, 2007; Mandel et al, 2010**), FcGR3A (**Breunis et al, 2009**) and FcGR3B (**Wilcocks et al 2008**). This gene dosage effect is

believed to expose CNV to evolutionary pressures (**Zhang et al, 2009**). From these observations it is thought that CNV can be a component of long term evolution, where the duplicated genes can develop new functions (“neofunctionalisation”). Example of such evolution can be found in the; immunity genes (**Hurles, 2004; Breunis et al, 2009**), olfactory genes (**Rouquier et al, 1998**) or opsin genes (**Nathans et al, 1986**).

It is generally unknown whether certain copy number “identities” (such as low or high copy number) are associated with haplotype blocks. Some CNV regions (CNVRs) have had coincidental analysis into haplotypes of CNV identities in the past. For example, the FcGR3 locus, which encodes a low affinity IgG receptor. One piece of evidence of possible haplotype structure at the locus is a systemic lupus erythematosus (SLE) associated SNP (FcGR2B-I232T). This was crucially found to be associated with a low CN of FcGR3B (**Niederer et al, 2010**), what the causative source of this association is, is unknown. Another piece of evidence is a study in 2008 (**Breunis et al**), which found 4 ‘patterns’ of CNV based on which three FcGR2/3 genes were duplicated as part of a tandem.

It is unknown whether these patterns are discrete “CNV haplotypes” that are being transmitted or coincidentally similar patterns of CNV, and serve as an example of the lack of general knowledge of how CNVR relate to the appearance of the haplotype structure of the human genome.

Another CNVR with a proposed evolutionary component, but no haplotype analysis, is the AMY1 gene locus (1p21 – **Entrez Gene**). The AMY1 gene encodes a member of the amylase gene family; secreted enzymes that hydrolyse α 1,4-glucoside bonds in starch to produce maltose, maltotriose and larger oligosaccharides. This is a process that starts in the mouth and is finished in the small intestine (**Mandel et al 2010**), with separate salivary (AMY1) and pancreatic (AMY2A and B) isoforms (**Samuelson et al, 1988**). Salivary amylase is a key component within saliva, being the most predominant protein in human saliva (**Oppenheim et al, 2007.**)

123 A structure of the AMY gene locus is shown in **FIGURE 1A**, as previously described
124 (**Samuelson et al, 1990**). The AMY1 and AMY2 genes share remarkably similar sequences,
125 showing as much as 98% sequence identity in their coding regions (**Horii et al, 1987**). Each
126 AMY gene contains a primate specific γ -actin pseudogene insert (**Samuelson et al, 1996**),
127 and the AMY1 specific retroviral insert into the pseudogene activates a cryptic promoter for

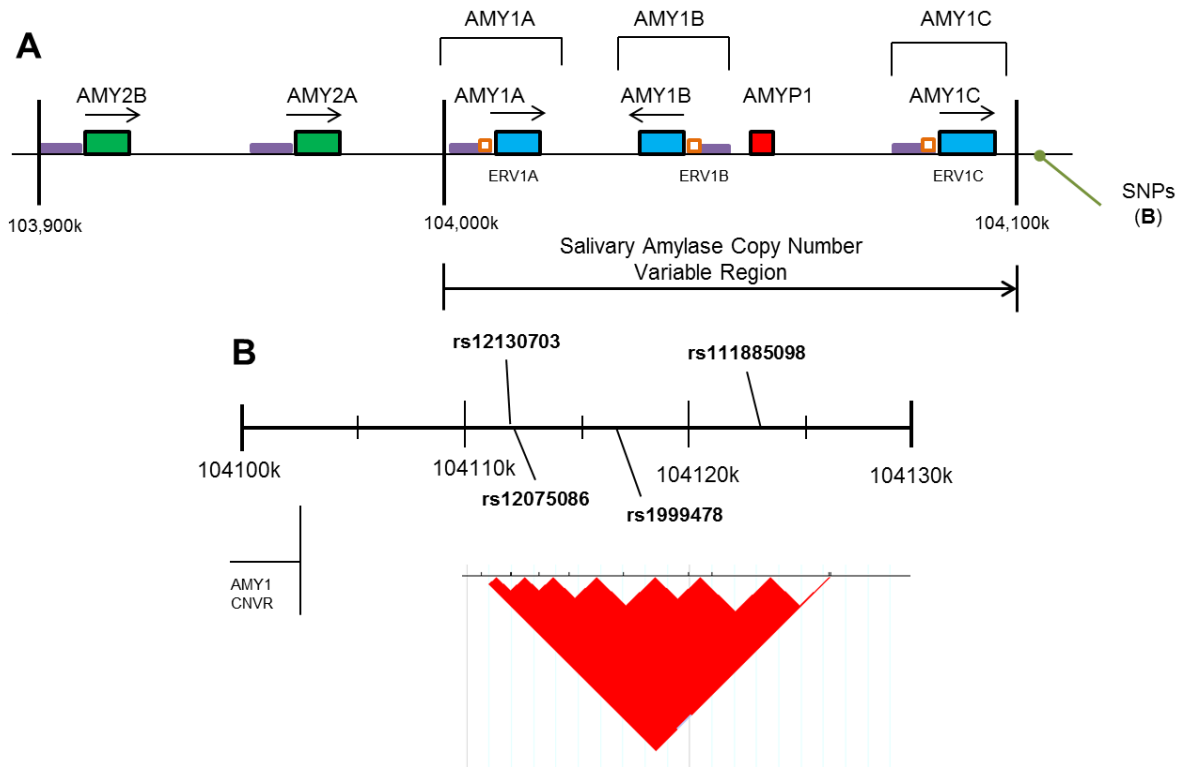


Figure 1

The AMY locus, on the short arm of chromosome 1. The positions of the AMY genes are shown (A), adapted from (**Samuelson et al, 1990**) with the copy number variable region indicated. AMY1 (salivary) are shown in blue with AMY2 (pancreatic) genes shown in green. The primate specific γ -actin pseudogene insert present at all amylase genes are shown in purple, with the AMY1 specific retroviral insert (ERV) shown in orange. This retroviral insert confers salivary specific expression (**Meisler and Ting, 1993**) by activating a cryptic promoter site in the γ -actin pseudogene. Arrows indicate gene direction, with AMY1B being inverted, and AMYP1 being a pseudogene. Also shown are the relative positions of the SNPs assayed (B), centromeric to the CNVR, within a block of linkage disequilibrium (shown by below LD plot, red indicates strong LD).

salivary expression. Studies have proven that this insert can cause transgenic salivary expression (**Meisler and Ting, 1993**).

CNV at the salivary amylase gene locus has been known for some time, being first described in 1989 (**Groot et al**). This study investigated the cause of previously described polymorphic salivary amylase expression patterns (**Pronk et al, 1982**). Restriction mapping demonstrated a CNV locus in a family pedigree, with two well described haplotypes, one with a 1 CN and another with 7. The paper concluded that this variation was due to homologous-unequal crossover, a model subsequently agreed on (**Samuelson et al, 1990**).

Salivary amylase protein levels are also shown to be polymorphic, and that these levels correlate with AMY1 CN ($r^2=0.351$), as described in 2007 (**Perry et al**). In their study they found that the correlation was very significant, with a $p < 0.001$ although some of the variation could not be explained just by CN. Subsequent findings have found this relationship to affect resultant activities of the action of salivary amylase. A study in 2010 (**Mandel et al**) found that higher salivary amylase, which their own data correlated with AMY1 CN, correlated with increased digestion of oral starch. A following study (**Mandel et al, 2012**) found that high AMY1 CN also correlated with an improved glycaemic homeostatic response to ingesting a starch solution. It is argued, therefore, that salivary amylase improved the anticipatory insulin response to carbohydrate ingestion, by an (at time of writing) unknown mechanism,

These finding suggest that the AMY1 CNV can be influenced by selective pressures. This argument is further supported by additional findings from Perry et al (**2007**). They described human populations that have had historically low starch consumption (classically a hunter-gather society), or have had historically high starch consumption (classically agricultural) varied in average CN. They reported that the low starch populations carried a low diploid CN (~4-5) whereas high starch population would have a high diploid CN (~6-9).

These observations build a model describing the variety of AMY1 CNV, that the increase of starch consumption has selected for higher CNs; as a high CN increases the ability of starch breakdown (**Mandel et al, 2010**), as well as improve the homeostatic response to

starch ingestion (**Mandel et al, 2012**). Varying population needs for starch intake would produce an overall variation in CN.

This model makes the salivary amylase CNVR an interesting target for evolutionary and population study. One way to investigate this is to ask whether CN identities are associated with surrounding haplotype structure. If so, then discrete patterns of CNV could be able to be tracked by studying a population's haplotype structure.

Therefore this study aims to genotype SNPs (and thereby capture haplotype block diversity) in several hundred individuals that have had their CNV also genotyped with the AMY1 locus. This CNV will have previously been genotyped, primarily by genotyping a microsatellite found at the locus. We will then compare the two data sets in each individual and ascertain whether any association with CNV and haplotype diversity can be found by statistically analysing the comparison.

Methods

rs12130703 and rs12075086 PCR

This was to amplify the PCR product that contained both rs12130703 and rs12075086 sites so that they can be later digested (by *BsrI* and *MnII* respectively), and was carried out in LD buffer. To a 20µl total; 10ng/µl DNA, 10µM forward and reverse primers, 5U/ µl Taq polymerase (NEB), 50 µM Tris-HCl pH8.8, 12.5mM Ammonium sulphate, 1.4mM MgCl₂, 7.5mM 2-mercaptoethanol, 200µM dNTPs and 125 µg/ml BSA. This produces a 1070 bp PCR product. This was amplified for 37 cycles, with each cycle being; 95° for 30 seconds, 56.8° for 30 seconds and 64.4° for 4 minutes, with a final extension step of 70° for 2 minutes. The forward primer was 5'GGTACAAACGAAAACACTGACA3' and the reverse primer was 5'TTGCTTACAATGCTTGACTTC3'. In order to develop this assay, known genotype individuals were used ECACC individual 39.

rs11185098 PCR

This was used to amplify the region containing rs11185098 so that it could be later digested by *Fnu4HI*. The PCR was carried out in ×10 buffer. To a 20µl total; 10ng/µl DNA, 10µM forward and reverse primers, 5U/ µl Taq polymerase (NEB), 50 µM Tris-HCl pH8.8,

12mM Ammonium sulphate, 5mM MgCl₂, 7.4mM 2-mercaptoethanol, 1.1mM dNTPs and 125 µg/ml BSA. This produced a 1400 bp product. This buffer solution is, compared to the LD buffer, enriched in dNTPs and MgCl₂ so that reactions, which produced low yield, produced enough for a restriction assay. This was amplified with an initial denaturing step of 95° for 30 seconds followed by 37 cycles of 95° for 30 seconds, 55.6° for 30 seconds and 65° for 3 minutes. The forward primer was 5'TTCAAAGATCCCCTTCCTT3' and the reverse primer was 5'ATTTTGGTGGCATTTCCTT3'. In order to develop this assay ECACC individual 39 was used.

Microsatellite PCR

This was used to amplify the tetra-repeat microsatellite believed to co-segregate with AMY1. It amplifies from the flanks of the repeat and therefore amplifies a product variable in size to the length of the amplified microsatellite. It was carried out in LD buffer, in a 10µl total; 10ng/µl DNA, 10µM forward and reverse primers, 5U/ µl Taq polymerase (NEB), 50 µM Tris-HCl pH8.8, 12.5mM Ammonium sulphate, 1.4mM MgCl₂, 7.5mM 2-mercaptoethanol, 200µM dNTPs and 125 µg/ml BSA. This was amplified with an initial denaturing step of 95° for 5 minutes. This was followed by 24 cycles of 95° for seconds, 53.6 for 30 seconds and 70° for 1 minutes, followed by a final extension step of 72° for 40 minutes. This final step was to make certain that the *Taq* added the A/T overhang to all products, as the microsatellite electrophoresis was sensitive to such base pair differences in length. The forward primer was FAM-5'ATTATCCTTTCACAGACAAAAG3' where FAM is a fluorescein derivative and creates a fluorescent primer, and the reverse was 5'TCCTCTAGGGTCATTCATTT3'.

Agarose Gel Electrophoresis

Gel electrophoresis was used to check the rs12130703, rs12075086 and rs11185098 PCRs success, as well as to visualise the RFLP assay. It was run in a 2% agarose gel of ×0.5 TBE buffer and 0.5µg/ml ethidium bromide. The 100 bp ladder was used for the *BsrI/MnII* gel runs, and a 1kb ladder was used for the *Fnu4HI* gel runs, because of the sizes of the PCR and RFLP products. These were carried out at 120V.

Restriction Fragment Length Polymorphism Assay

Bsrl Assay: For a 15µl total; 5µl of *Bsrl* PCR product, 1× NEB buffer 3, 2.5U of *Bsrl* and incubate overnight at 65°. **MnII assay:** For a 15µl total; 6µl of *MnII* PCR product, 1× NEB buffer 4, 2.5U of *MnII* and incubate overnight at 37°. **Fnu4HI assay:** For a 15µl total; 5µl of *Fnu4HI* PCR product, 1× NEB buffer 4, 1.5U of *Fnu4HI* and incubate overnight at 37°. NEB buffer 3 contains; 100mM NaCl, 50mM Tris-HCl, 10mM MgCl₂, 1mM dithiothreitol. pH7.9 @ 25°. NEB buffer 4 contains; 50mM Potassium acetate, 20mM Tris-acetate, 10mM Magnesium acetate, 1mM dithiothreitol. pH.7.9 at 25°.

To develop the *Bsrl* and *MnII* assays the following HapMap CEPH (Centre d'Etude du Polymorphisme Humain) plate 01 individuals were used as known genotype controls; NA12144, A12717, NA 12264, NA07357, NA11831, NA11830. To develop the *Fnu4HI* assay individuals from HapMap CEPH plat 01; NA12044, NA12815, NA12043, NA12892 and NA12873 were used.

Microsatellite Capillary Electrophoresis

For every 16 samples, 10µl was added to each sample from an initial mix of 2µl Rox molecular marker and 170µl HiDi formamide. 2µl of the microsatellite PCR product was then added to this. It was denatured at 95° for 3 minutes, then run on an ABI 3130xl (Applied Biosystems), with 1kV for 30 seconds.

Results

In order to initially assess the viability of investigating whether haplotypes can be used to predict copy number identity, preliminary work was first carried out on existing data with individuals from the International HapMap Project, CEPH plate 01. These individuals are well described in terms of SNPs, allowing haplotypes to be inferred from tagging SNPs.

Individuals from the HapMap project were useful because the individuals described are described as haploid individuals. This is a result of being part of well-defined pedigrees. Following the transmission of haplotypes through these families allows the segregation of diploid genotypes into haplotypes. This “segregation analysis” allowed us to directly compare

the CNV haplotype to the microsatellite haplotypes, which increases the statistical power of the preliminary work.

These “HapMap” individuals had previously had their copy number genotyped by a combination of a microsatellite genotype assay and AMY1 CNV PRT and non-PRT tests. This “microsatellite assay” is a PCR amplification of all copies of that microsatellite, which are then separated by size via capillary electrophoresis, yielding a genotype based on the length of each repeat, and the amount of repeats. This assay was carried out by Sugandha Dhar, a PhD student within the lab.

Therefore in order to compare haplotype to CNV, only capturing the haplotype diversity was needed. Consequently, these CEPH individuals’ SNP genotypes were investigated using the HapMap genome browser, release #24. Analysis into the SNP diversity at adjacent (centromeric) LD blocks (chr1:104110000-104130000) yielded three SNPs, rs12130703, rs12075086 and rs1999478, see **FIGURE 1B** for positions. These segregated the data into a reasonable amount of haplotypes with which to compare to CNV. These SNPs would therefore tag the representative haplotypes within the European population samples used, as most haplotype blocks have 2-4 main haplotypes presented within the population (**Daly et al, 2001**). These chosen three SNPs tagged the segregation of individuals into 5 distinct haplotypes, designated; 1, 2, 3, 4 and 5.

After sorting individuals into their component haplotypes, associations with CNV were investigated. In the preliminary data, which had a sample size of 46, it appeared that these SNP haplotypes did not have any significant associations with total CN. However there did appear to be association with haplotypes 3 and 5 and an enrichment of short microsatellites (lengths <265 bp) (Chi-squared value = 7.30, $p = 0.007$, $n = 46$). This suggested that any association with haplotype structure and CNV identity would be found in microsatellite length variation, not total CN. However a larger data set was needed to confirm these hypotheses, and provide more analytical power. The larger data set would also be used to confirm the lack of association with CN, which was particularly surprising.

The three SNPs (rs12130703, rs12075086 and rs1999478) were then genotyped in ECACC individuals. These are “European Collection of Cell Cultures” individuals and were an independent data set from the HapMap individuals, allowing confirmation of the preliminary observations. These individuals had already been genotyped with respect to the AMY1 microsatellite by Sugandha Dhar, so that only the SNP genotyping needed to be carried out.

In order to be able to genotype the SNPs, restriction fragment length polymorphism (RFLP) assays were developed for the three SNPs chosen. The rs12130703 and rs12075086 polymorphisms were found to create or destroy a *BsrI* and *MnII* site, respectively. The rs1999478 polymorphism did not create a restriction site (sequence; CCCCNAAAA, where n is rs1999478 – a C/A polymorphism) and the surrounding sequence made an allele specific PCR assay unsuitable for high-throughput genotyping. Therefore rs1999478 was abandoned as a testable SNP.

This meant, however, that haplotypes 1/3, and 2/4 could not be distinguished as the identity of rs1999478 segregated 1/3, and 2/4. Therefore another SNP was sought to either “tag” rs1999478’s segregation, or a SNP that would segregate the individuals in a similar manner, keeping as much of the original haplotypes as possible.

A new SNP, rs11185098, was found, preserving the old haplotypes, with the exception of 1 and 3. Under segregation by rs11185098, haplotype 3 gains some individuals that previously belonged to haplotype 1, the remaining forming the new haplotype 1. In order to genotype rs11185098 the enzyme *Fnu4H1* is used in a RFLP assay, similarly to rs12130703 or rs12075086. **FIGURE 2** shows representative digestions of the *BsrI*, *MnII* and *Fnu4H1* assays.

TABLE 1 demonstrates the preliminary data, split by the SNPs; rs12130703, rs12075086 and rs11185098.

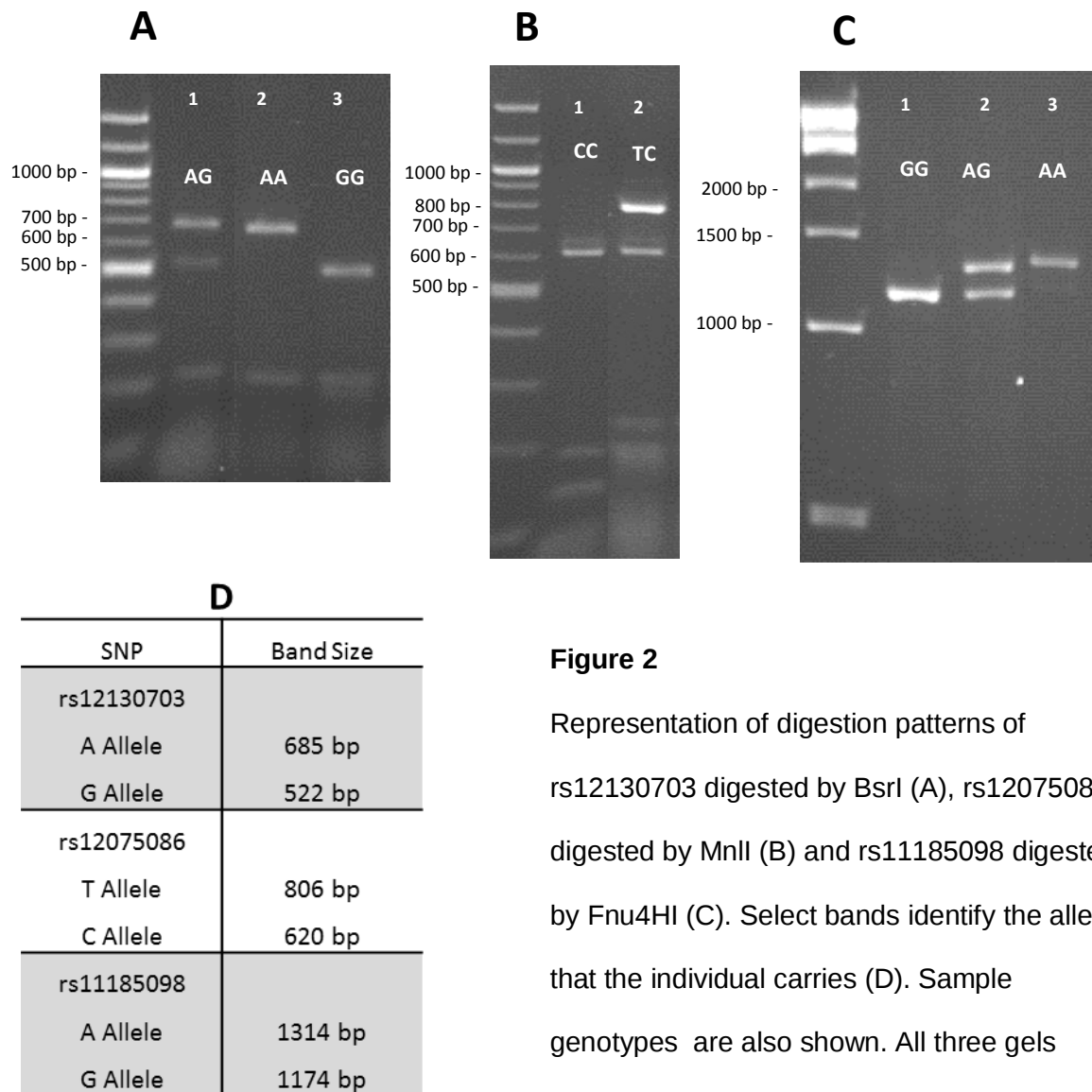


Figure 2

Representation of digestion patterns of rs12130703 digested by BsrI (A), rs12075086 digested by MnlI (B) and rs11185098 digested by Fnu4HI (C). Select bands identify the alleles that the individual carries (D). Sample genotypes are also shown. All three gels represent different individuals.

The data produced in this phase of genotyping produced diploid genotypes. This is due to the unrelated nature of the ECACC individuals, and therefore the SNP haplotype and CNV haplotype were not able to be easily phased (unlike in the case of the HapMap individuals).

The diploid manner of the ECACC individuals did not cause the phasing of the SNPs to be impossible, however. The known associated nature of SNP to SNP allowed certain assumptions to be made in order to phase the diploid genotype (Clark, 1990). For example knowing that allele rs12130703A is only associated with haplotype 5, allows AA and AG

CNV and Haplotype Association within the AMY1 locus.
Edmund Gilbert

299

Table 1										
The results of the preliminary investigation. Shown are the HapMap individuals tested; their haplotype as defined by rs12130703, rs12075086 and rs11185098 as well as their microsatellite (MS) profile, and total copy number (CN). The individuals are defined as their HapMap designated reference, as well as the chromosome the haplotype originated on.										
Haplotype	SNP Identity			HapMap Ref	Microsatellite Identity					Average Haplotype CN
	rs12130703	rs12075086	rs11185098		MS Length and CN				CN Total	
1	G	C	A	NA12236_c1			1		1	3.40
				NA11832_c1		1		1	2	
				NA11832_c2				3	3	
				NA12145_c1			2	2	4	
				NA12146_c1			3	3	1	
2	G	T	A	NA12155_c2		1	1		2	3.88
				NA12043_c2		1	5		6	
				NA12249_c2		1	2	2	5	
				NA12264_c2		1	2	1	1	
				NA11994_c2				1	1	
				NA12239_c2			1		1	
				NA11831_c1		1		3	4	
				NA11830_c1		1	3	2	6	
3	G	C	G	NA07034_c1			2		2	2.76
				NA12156_c1			2		2	
				NA12043_c1		2	4		6	
				NA06993_c2			3		3	
				NA07357_c1			3		3	
				NA12264_c1			3		3	
				NA07055_c1		1	1	1	3	
				NA07357_c1		2	1		3	
				NA12145_c2			1	1	2	
				NA12239_c1		1	2		3	
				NA11993_c1		1	2		3	
				NA11829_c1		3			3	
				NA07345_c1			4		4	
				NA12155_c1			1		1	
				NA12154_c2			2		2	
				NA12234_c2	1		2		3	
				NA12249_c1		1	2		3	
				NA12154_c1	2		1		3	
				NA11994_c1	1				1	
				NA11829_c2			2		2	
				NA12146_c2			3		3	
4	G	T	G	NA12236_c2			3		3	3.00
5	A	C	G	NA12248_c2			1	1	2	3.73
				NA12044_c1				3	3	
				NA12044_c1	1		1	1	3	
				NA06993_c1			2		1	
				NA12156_c2		2	1		3	
				NA12144_c2			2		2	
				NA12144_c1		1	2		3	
				NA11830_c2	2		1	1	4	
				NA11992_c1		2	1	2	5	
				NA11993_c2			4	1	5	
				NA11831_c2			5	3	8	

genotypes to be defined immediately as respectively 5|5 or 5|any individuals. The diploid SNP genotypes that correspond to the diploid combinations of the haplotypes are shown in **TABLE 2**. SNP genotype and CNV genotype were unable to be phased, however. This meant that analysis would have to take into account the homozygous and heterozygous nature that we would see in a diploid population. The Cochran-Armitage Test (**Cochran, 1954**) (**Armitage, 1955**) would be able to test such a system and rescue some of the lost power that analysing diploid data would cause. This test is a variant of the Chi-Squared Test, and tests for an association that is 'dose' dependent, and can be used in genetics to test for homozygote (++), heterozygote (+-) and homozygote (--) differences.

The direct output of the microsatellite only gives the ratios of microsatellite length copy number (see **FIGURE 3**), so the majority of analysis was based on the presence/absence of microsatellite lengths, not total copy number. This was not deemed a loss of significant data as the preliminary data presented no clear evidence of association of SNP haplotype with total CN.

Table 2			
The diploid SNP genotypes of all the permutations of the 5 haplotypes identified, allowing the genotyping of individuals based on their SNP diploid genotypes.			
Diploptype	rs12130703	rs12075086	rs11185098
1 1	GG	CC	AA
1 2	GG	CT	AA
1 3	GG	CC	AG
1 4	GG	CT	AG
1 5	AG	CC	AG
2 2	GG	TT	AA
2 3	GG	CT	AG
2 4	GG	TT	AG
2 5	AG	CT	AG
3 3	GG	CC	GG
3 4	GG	CT	GG
3 5	AG	CC	GG
4 4	GG	TT	GG
4 5	AG	CT	GG
5 5	AA	CC	GG

CNV and Haplotype Association within the AMY1 locus.
Edmund Gilbert

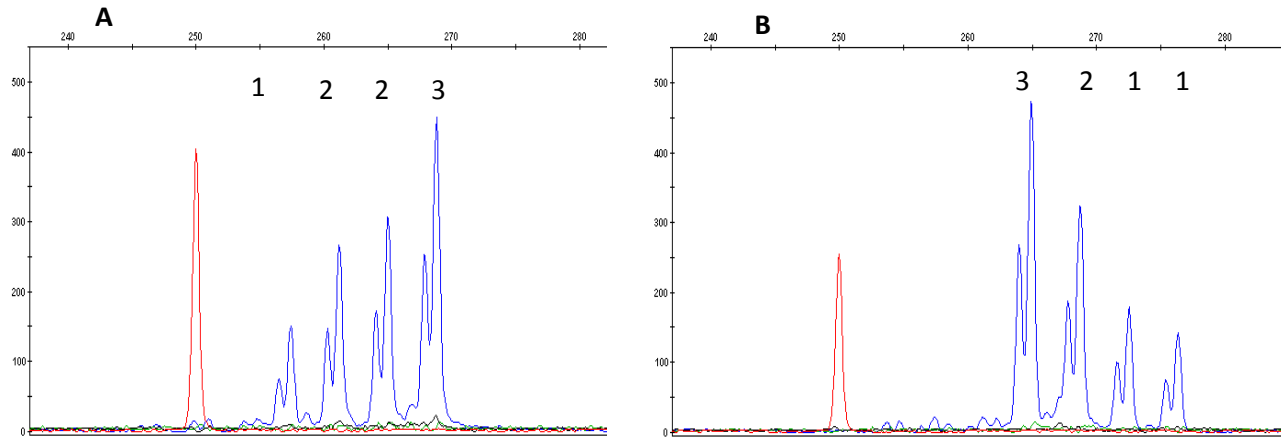


Figure 3

Representative microsatellite trace, showing relative microsatellite intensities. Higher relative peak height (y axis) indicates a higher copy number at that microsatellite size (x axis) relative to lower peaks. Therefore (A) shows a 1:2:2:3 ratio (with respect to the sizes; 257, 261, 265 and 269 bp). This corresponds to either an 8 copy individual, or a 16 (which would produce the same ratio pattern); however most likely the individual is an 8 copy, as 16 copy individuals are rarer (Perry et al, 2007). The red peaks shown are molecular markers, of a constant intensity of 400 and size of 50 bp intervals. Data shown is my own, not of Sugandha Dhar's wider genotyping work.

From this stage of the experiment 450 individuals were initially genotyped, with respect to the SNPs rs12130703, rs12075086 and rs11185098. Due to some PCR/RFLP reactions not working and not all individuals were able to be retyped, this study uses the 370 individual that were able to be defined. This data set forms the basis of the following analyses, and a summary of the SNP identity, and frequency, of the haplotypes are shown in **TABLE 3**.

With the data set gathered, before the testing of association hypotheses, it was important to ascertain whether the allele calling and the assumptions of SNP

Table 3				
The five haplotypes described in this study, and the SNP identity that tags the haplotypes. Also shown are the frequencies of the 5 haplotypes within the ECAC individuals tested.				
Haplotype	rs12130703	rs12075086	rs11185098	Frequency
1	G	C	A	0.137
2	G	T	A	0.052
3	G	C	G	0.397
4	G	T	G	0.001
5	A	C	G	0.414

association produced a dataset that could be trusted. The genotype frequencies of each

SNP were tested against an expected genotype frequencies, based on the allele frequencies, under the Hardy-Weinburg Equilibrium (HWE). The *BsrI* and *MnII* assays presented no significant departure from observed and expected (Chi squared $p = 1.000$, Chi squared $p = 0.979$ respectively). The *Fnu4HI* assay however presented a significant departure from the expected (Chi squared $p = 0.04$). This resulted from an under calling of the AA allele, and after independent corroboration of genotypes the problem continued. We do not believe that this under-calling is a result of systematic error, however we cannot account for the lack of AA individuals. The association between the three SNPs being genotyped was also checked, and the results are shown in **TABLE 4**. The D' value of 1 indicates the SNPs being in strong association, where we do not see all the possible genotypes. The r^2 value indicates association, but not absolute, which is to be expected as the three vary enough to tag the partitioning of the haplotypes.

Within the preliminary data it was observed that haplotype 3 contained a null 269 microsatellite length alleles that was associated with no

Table 4				
The results of testing the association strength and confidence of the three SNPs genotyped within the study.				
SNPs Compared		D' Value	r2 Value	p Value
rs12130703	rs12075086	1	0.060	1.56E-11
rs12075086	rs11185098	1	0.38	2.21E-66
rs12130703	rs11185098	1	0.16	3.22E-28

microsatellite lengths >269. This prompted the hypothesis whether this was indicative a subpopulation of haplotype 3; if it would be we'd expect a bias within haplotype 3 for more null 269 alleles. Therefore, whether there was any significant difference between the numbers of null 269 alleles in haplotype 3 or other haplotypes in the larger data set was investigated. However it was found that there was no significant difference ($Z = 1.141$, $p = 0.127$, $n = 376$). The observations that haplotypes 3 and 5 were found to be associated with short alleles (lengths of 257 and 261) proved to be insignificant within the large data set ($Z = 0.724$, $p = 0.235$, $n = 376$); as did the same relationship with haplotypes 3 and 5 alone ($Z = 0.326$, $p = 0.372$, $n = 376$) ($Z = 0.929$, $p = 0.177$, $n = 376$), respectively.

The ratio nature of the microsatellite data meant that the true integer CN value for the ECACC individuals could not be obtained. However we could round up the ratio values to

integers in order to produce some CN values with which to test. We used a Cochran-Armitage test defining case/control as those less than or equal to or greater than the total population median (6); haplotype 1 ($Z = 1.28$), haplotype 2 ($Z = 0.40$) and haplotype 5 ($Z = 1.06$) with all $p = >0.1$ and $n = 370$. Haplotype 3 appeared to have a significant association ($Z = 1.67$, $p = 0.05$, $n = 370$), but further analysis demonstrated that this was not truly significant.

Discussion

This study has found that there does not seem to be any significantly strong correlation with haplotype structure and CNV within the AMY1 locus. This is initially surprisingly, given a literature background that suggests that the locus has had a significant evolutionary history within the human population (**Perry et al, 2007; Mandel et al, 2010 and 2012**).

The relatively small sample size of 370 means that we would not have the power to observe any subtle associations that may exist. The lack of strong association suggests two possibilities. Firstly, the sequence between SNPs and the CNVR are being recombined, so that the two are not linked. Or that the generation of new CNV identities is faster relative to the generation of new SNP haplotypes, disrupting any association. Considering the extent of high LD around the CNVR when we study the region in the HapMap genome browser, we believe that the latter explanation is more likely.

The microsatellite information produces minimum ratios (see **FIGURE 3**), and not total copy number. Nevertheless we could still use this data to investigate CN as the minimum ratio (and resultant total CN) is still based on the 'true' CN. One issue is of multiples; a 1:2:1 ratio could be produced by a 4 copy individual or an 8 or 12 copy individual. This 'rounding down' to the minimum CN does cause the data set used to have a systematic bias to calling such individuals as lower CN than they would otherwise be called. This would mean that though in the larger picture CN would broadly reflect the population's CN distribution, the observed average CN will be called lower than it otherwise would. This means that the significant association with haplotype 3 we have observed ($Z = 1.67$, $p = 0.05$, $n = 370$) is most likely insignificant if we imagine the removal of a bias towards lower CN. Work using

non-PRT and PRT tests in conjunction with the microsatellite data to produce a true CN profile for each individual, via a maximum-likelihood method, as (as of yet) be completed within the laboratory by Sugandha Dhar. For this reason the use of microsatellite ratios had to be used.

Another issue with using the rounded ratios of microsatellites is the idea of a 'minimum ratio value'. If a ratio gives 1:2:3 we could be quite confident that the minimum ratio value would be 1. If we imagine a situation where the ratios gave 1:1.5:1.5 we may round up to 1:2:2 to get a total of 5 and a minimum ratio value of 1. However, the ratio pattern 1:1.5:1.5 could easily represent a 2:3:3 ratio (and now a 8 copy individual); the process to convert in ratio having reduced this to a minimum ratio value of 1. Therefore analysis into whether individuals with ratio numbers around $n.5$ at one or more lengths. 15 such individuals identified were. These individuals' genotypes were then converted into CNV identities that better represented the microsatellite ratio implications and incorporated into the final analyses.

We have also assumed that every copy of the AMY1 gene is completely associated with 1 copy of the microsatellite. Unpublished data suggests that this is the case, with no known exceptions. If this were the case however, the CN would only vary slightly, and therefore not significantly impact the study.

The haplotypes studied are not homogenous in their diversity within the population. The most frequent of the haplotypes were 3 and 5, being almost equally prominent (see **TABLE 3**). This prominence was so much so that haplotypes 3 and 5 represent 80% of the population. These frequencies support previous literature on the limited diversity within haplotype blocks (**Daly et al, 2001; Patil et al, 2001**), increasing confidence that we have captured the representative haplotype diversity at this locus.

The method of calling the diploid haplotype genotypes resulted in the majority of expected SNP diploid genotypes being able to be defined, but did assume no new haplotypes were observed. We did not observe any genotypes unable to be explained by

TABLE 2, the only anomalous genotype (data not shown) was found to be a genotyping error. Therefore the rate of false positives is believed to be insignificant.

The only ambiguous diploid genotype was GG/CT/AG (rs12130703, rs12075086 and rs11185098 respectively). This meant that the 1|4 and 2|3 diploid genotypes were indistinguishable. However these individuals were all used in later analysis, as 2|3 individuals. This was because the rarity of haplotype 4 (being observed only once in this study) meant that of the total 35 ambiguous individuals, no haplotype 4 would be reasonably expected. Whilst haplotype 4's rarity does not contribute to the studies focus on association it does help increase confidence in the accuracy of our capture of haplotype diversity; in that we are observing the rare haplotypes expected in a population study.

The SNPs that were used to tag the haplotype structure at the AMY1 locus were all significantly associated with each other, see **TABLE 4**. The p values for the D' and r^2 values are all extremely significant and provide us confidence that these SNPs tag LD structure around the AMY1 locus. The D' (**Hedrick, 1987**) values for all the pairs are 1. Given that not all the possible four genotypes have been observed this is an unsurprising value. The r^2 values are somewhat lower than the D' values, this was also expected because each pair is not associated with only one genotype (which would return an r^2 value of 1). The variation at these loci that produces this r^2 value was the reason for those SNPs to be chosen, so that they could tag multiple haplotypes.

This study also concentrated on European samples, both in the preliminary data that used samples from CEPH and testing a wider population size within ECACC individuals. The European sample group was felt to provide enough variation to investigate whether there was any general association with haplotypes and CNV. However we would expect to observe other haplotypes in other populations, such as Asian or African. With the latter we would expect a greater extent of variation (**Frisse et al, 2001**), so were not chosen for such a preliminary study as the greater variability would obscure any association we would or would not expect to see.

The expectation that the European samples represent a subdivision within the greater human diversity is supported by wider SNP diversity within the haplotypes tested. General SNP background of the haplotypes indicate that haplotype 5 is part of an outlier group of haplotypes; with the 5/other split represented by a clear division in general SNP diversity. This segregation can be defined (within the haplotypes studied) by the possession of an A allele at rs12130703. Cochran-Armitage analysis of CN between this 5 versus the rest also resulted in a p value >0.05. Perhaps investigation within other populations would provide other, intermediate haplotype blocks.

It is entirely reasonable that given a strong selective force for certain AMY1 CNs in a recent time and/or small isolated population, that an association with haplotype and CNV could be observed. Such studies would be very interesting to see if there are any populations with such strong evidence for recent selection in AMY1 considering the previous literature (**Perry et al, 2007; Mandel et al, 2010 and 2012**) and the lack of association found in this study.

We have only seen the 4 major haplotypes of the locus; 1, 2 3 and 5 with haplotype 4 being a rare haplotype. We did not observe any other haplotypes than the ones that were observed in the preliminary stage of investigation. Therefore the 4 major haplotypes we see in this large data set represent the majority of European individual haplotypes. We would expect that in a larger scale investigation more rare haplotypes, such as haplotype 4, would be observed. However we would expect that these haplotypes are not relevant to the question of if the AMY1 locus CNV is associated with haplotype structure. Considering the lack of association structure with the major haplotypes of the locus, the rare haplotypes would not be expected to have any association structure, as well as their rarity would present analytical difficulty.

The four SNPs that have been discussed (rs12130703, rs12075086, rs11185098 and the abandoned rs1999478) are not the only SNPs at the loci studied – the centromeric and telomeric adjacent LD blocks. Most of the SNPs studied in European samples at the centromeric locus segregate at the major two haplotype split, which is tagged by

rs12130703. A handful of SNPs were observed to vary enough, these would segregate the haplotypes into more defined partitions. The SNPs investigated (such as rs12130703, rs12075086, rs1999478 and rs11185098) were these SNPS. Also, the aim of this study was to compare the general haplotype structure with AMY1 CNV and to observe whether there was any general association, not to provide a lineage map of the adjacent haplotypes. Considering that the centromeric haplotype block has no association with the CNV it is unlikely that the telomeric haplotypes would be associated with CNV, so the SNPs that are within that block do not need to be considered for the initial hypothesis of this study.

The SNP allele frequencies are in agreement to the HapMap frequencies (all $p = >0.9$) enough to be confident that the frequencies of the allele, genotype and haplotype are a true representation. The sample sizes of the HapMap derived allele frequencies cluster around 100 individuals, so this larger data set may be expected to not match perfectly because the difference in random sampling variation.

We also tested whether the observed genotype frequencies were significantly different than the expected genotype frequencies derived from the HWE ($pp+2pq+qq$), if there was a departure from the HWE. The observed versus expected for rs12130703 and rs12075086 were not-significant ($p = >0.95$), however there was a significant departure in the case of rs11185098 ($p = 0.04$). This significance is removed if the amount of AA genotype calls were to be increased by about 7. It seems, therefore, the genotype AA is has been under-called in the study, and one explanation is that the calling process somehow had a systematic bias for calling AA another genotype. However, upon closer inspection were the entire data set was re-called independently of the original caller this problem still arose. Following such analysis we are confident that the genotypes that have been called at the rs11185098 position are correct, however we have not been able to account for this under-representation of the AA genotype.

In conclusion, this study has found no significant association of CNV with haplotype structure; we believe that this is because the generation of CNV at the AMY1 locus is faster

than the generation of new haplotypes. This would therefore disrupt any association that appeared historically. We believe further research into the extent of this lack of association, such as into isolated populations can shed light on the evolutionary structure of copy number variation at the AMY1 locus.

Acknowledgements

We would like to acknowledge; Victoria King for her help on designing the rs12130703 and rs1207035086 assays, Sugandha Dhar for her AMY1 microsatellite and PRT/non-PRT test data which allowed us to investigate CNV, and John Armour and Jess Tyson for their invaluable help and advice.

References

1. Armitage P. (1955) Tests for Linear Trends in Proportions and Frequencies. *Biometrics*. 11: 375–386.
2. Armour J, Palla R, Zeeuman P, Heijer M, Schalkwijk J and Hollox E. (2007) Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res*. 35: e19.
3. Bailey J, Liu G and Eichler E. (2003) An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *Am J Hum Genet*. 73: 823-34.
4. Bailey J, Kidd J and Eichler E. (2009) Human copy number polymorphic genes. *Cytogenet Genome Res*. 123: 234-243.
5. Breunis W, van Mirre E, Bruin M, Geissler J, de Boer M, Peters M et al. (2008) Copy number variation of the activating FCGR2C gene predisposes to idiopathic thrombocytopenic purpura. *Blood*. 111: 1029-38.
6. Clark A. (1990) Inference of Haplotypes from PCR-amplified Samples of Diploid Populations. *Mol Biol Evol*. 7: 111-122.
7. Cochran WG. (1954) Some methods for strengthening the common chi-squared tests. *Biometrics*. 10: 417–451.
8. Conrad D, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y et al. (2010) Origins and functional impact of copy number variation in the human genome. *Nature*. 464: 704-12.
9. Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D and Scozzari R. (2011) A Revised Root for the Human Y Chromosomal Phylogenetic Tree: The Origin of Patrilineal Diversity in Africa. *Am J Hum Genet*. 88: 814-818.
10. Daly M, Rioux J, Schaffner, Hudson T and Lander E. (2001) High-resolution haplotype structure in the human genome. *Nat Genet*. 29: 229-32.
11. Frisse L, Hudson R, Bartoszewicz A, Wall J, Donfack J and Di Rienzo A. (2001) Gene Conversion and Different Population Histories May Explain the Contrast between Polymorphism and Linkage Disequilibrium Levels. *Am J Hum Genet*. 69: 831-43.
12. Gabriel S, Schaffner S, Nguyen H, Moore J, Roy J, Blumensteil B et al. (2002) The Structure of Haplotype Blocks in the Human Genome. *Science*. 296: 2225-9.
13. Groot P, Bleeker M, Pronk J, Arwert F, Mager W, Plana R et al. (1989) The human α -amylase multigene family consists of haplotypes with variable numbers of genes. *Genomics*. 5: 29-42.
14. Hastings P, Lupski J, Rosenberg S, Ira G. (2009) Mechanisms of change in gene copy number. *Nat Rev Genet*. 10: 551-64.
15. Hedrick P. (1987) Gametic Disequilibrium Measures: Proceed With Caution. *Genetics*. 117: 331-341.
16. Hollox E, Armour J and Barber J. (2003) Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *Am J Hum Genet*. 73: 591-600.
17. Horii A, Emi M, Tomita N, Nishide T, Ogawa M, Mon T and Matsubara M. (1987) Primary structure of human pancreatic c*-amylase gene: its comparison with human salivary a-amylase gene. *Gene*. 60: 57-64.
18. Hurles M. (2004) Gene Duplication: The Genomic Trade in Spare Parts. *Plos Biol*. 2: e206.
19. Jakobsson M, Scholz S, Scheet P, Gibbs R, VanLiere J, Fung H et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*. 451: 998-1003.
20. Jeffreys A, Kauppi L, and Neumann R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet*. 29: 217-222.
21. Johnson G, Esposito L, Barratt B, Smith A, Heward J, Di Genova G et al. (2001) Haplotype tagging for the identification of common disease genes. *Nat Genet*. 29: 233-7.

22. Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V et al. (2001) An Extensive Analysis of Y-Chromosomal Microsatellite Haplotypes in Globally Dispersed Human Populations. *Am J Hum Genet.* 68: 990-1018.
23. Kidd J, Cooper G, Donahue W, Hayden H, Sampas N, Graves T et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature.* 453: 56-64.
24. Koene H, Kleijer M, Roos D, De Haas M and Von dem Borne A. (1998) FcγRIIIB Gene Duplication: Evidence for Presence and Expression of Three Distinct FcγRIIIB Genes in NA(1+,2+)SH(+) Individuals. *Blood.* 91: 673-9.
25. Lewontin and Kojima. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution.* 14: 458-72.
26. Mandel A, Peyrot des Gachons C, Plank K, Alarcon S and Breslin P. (2010) Individual Differences in AMY1 Gene Copy Number, Salivary α-Amylase Levels, and the Perception of Oral Starch. *Plos One.* 5: e13352.
27. Mandel A and Breslin P. (2012) High Endogenous Salivary Amylase Activity Is Associated with Improved Glycemic Homeostasis following Starch Ingestion in Adults. *J Nutr.* 142: 853-8.
28. Meisler M and Ting C. (1993) The Remarkable Evolutionary History of the Human Amylase Genes. *Crit Rev Oral Biol Med.* 4: 503-9.
29. McCarroll S, Huett A, Kuballa P, Chilewski S, Landry and Goyette P et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet.* 40: 1107-12.
30. Nathans J, Piantanida T, Eddy R, Shows T and Hogness D. (1986) Molecular genetics of inherited variation in human color vision. *Science.* 232: 203-210.
31. Niederer H, Willcocks L, Rayner T, Yang W, Lau Y, Williams T et al. (2010) Copy number, linkage disequilibrium and disease association in the FCGR locus. *Hum Mol Genet.* 19: 3282-94.
32. Oppenheim F, Salih E, Siqueira W, Zhang W and Helmerhorst E. (2007) Salivary Proteome and Its Genetic Polymorphisms. *Ann N Y Acad Sci.* 1098: 22-50.
33. Patil N, Berno A, Hinds D, Barrett W, Doshi J, Hacker C et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science.* 294: 1719-23.
34. Perry G, Dominy N, Claw K, Lee A, Fiegler H, Redon R et al. (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet.* 39: 126-60.
35. Phillips M, Lawrence R, Sachidanandam R, Morris A, Balding D, Donaldson M et al. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet.* 33: 382-7.
36. Piotrowski A, Bruder C, Andersson R, Diaz de Ståhl T, Menzel U, Sandgren J et al. (2008) Somatic mosaicism for copy number variation in differentiated human tissues. *Hum Mutat.* 29: 1118-24.
37. Pronk J, Frants R, Jansen W, Eriksson A and Tonino G. (1982) Evidence for duplication of the human salivary amylase gene. *Hum Genet.* 60: 32-5.
38. Redon R, Ishikawa S, Fitch K, Feuk L, Perry G, Andrews D et al. (2009) Global variation in copy number in the human genome. *Nature.* 444: 444-4.
39. Rouquier S, Taviaux S, Trask B, Brand-Arpon V, van den Engh G, Demaille J et al. (1998) Distribution of olfactory receptor genes in the human genome. *Nat Genet.* 18: 243-250.
40. Samuelson L, Wiebauer K, Gumucio D and Meisler M. (1988) Expression of the human amylase genes: recent origin of a salivary amylase promoter from an actin pseudogene demonstrates the split between pancreatic and salivary. *Nucleic Acids Res.* 16: 8261-76
41. Samuelson L, Wiebauer K, Snow C and Meisler M. (1990) Retroviral and Pseudogene Insertion Sites Reveal the Lineage of Human Salivary and Pancreatic Amylase Genes from a Single Gene during Primate Evolution. *Mol Cell Biol* 10: 2513-20.

- 614 42. Samuelson L, Phillips R and Swanberg L. (1996) Amylase Gene Structures in
615 Primates: Retroposon Insertions and Promoter Evolution. *Mol Cell Evol.* 13: 767-79.
616 43. Slatkin M. (2008) Linkage disequilibrium — understanding the evolutionary past - and
617 mapping the medical future. *Nat Rev Genet.* 9: 477-85.
618 44. Smith J and Haigh J. (1974) The hitch-hiking effect of a favourable gene. *Genet Res.,*
619 *Camb.* 23: 23-35.
620 45. Underhill P, Shen P, Lin A, Jin L, Passarino G, Yang W et al. (2000) Y chromosome
621 sequence variation and the history of human populations. *Nat Genet.* 26: 358-61.
622 46. Wall J and Pritchard J. (2003) Haplotype blocks and linkage disequilibrium in the
623 human genome. *Nat Rev Genet.* 4: 587-97.
624 47. Wilcocks L, Lyons P, Clatworthy M, Robinson J, Yang W, Newland S et al. (2004)
625 Copy number of FCGR3B, which is associated with systemic lupus erythematosus,
626 correlates with protein expression and immune complex uptake. *J Exp Med.* 205:
627 1573-82.
628 48. Zegura S, Karafet T, Zhivotovsky L and Hammer M. (2004) High-Resolution SNPs
629 and Microsatellite Haplotypes Point to a Single, Recent Entry of Native American Y
630 Chromosomes into the Americas. *Mol Biol Evol.* 21: 164-75.
631 49. Zhang K, Calabrese P, Nordborg M and Sun F. (2002) Haplotype Block Structure
632 and Its Applications to Association Studies: Power and Study Designs. *Am J Hum*
633 *Genet.* 71: 1386-94.
634 50. Zhang K, Qin Z, Liu J, Chen T, Waterman M and Sun F. (2004) Haplotype Block
635 Partitioning and Tag SNP Selection Using Genotype Data and Their Applications to
636 Association Studies. *Genome Res.* 14: 908-16.
637 51. Zhang F, Gu W, Hurler M and Lupski J. (2009) Copy Number Variation in Human
638 Health, Disease, and Evolution. *Annu Rev Genom Hum Genet.* 10: 451-81.